

# NLP for Counterspeech against Hate: A Survey and *How-To* Guide

Helena Bonaldi<sup>1,2</sup> Yi-Ling Chung<sup>3</sup> Gavin Abercrombie<sup>4</sup> Marco Guerini<sup>1</sup>

<sup>1</sup>Fondazione Bruno Kessler, Italy, <sup>2</sup> University of Trento, Italy,

<sup>3</sup>The Alan Turing Institute, <sup>4</sup>The Interaction Lab, Heriot-Watt University

hbonaldi@fbk.eu, ychung@turing.ac.uk, g.abercrombie@hw.ac.uk, guerini@fbk.eu

## Abstract

In recent years, counterspeech has emerged as one of the most promising strategies to fight online hate. These non-escalatory responses tackle online abuse while preserving the freedom of speech of the users, and can have a tangible impact in reducing online and offline violence. Recently, there has been growing interest from the Natural Language Processing (NLP) community in addressing the challenges of analysing, collecting, classifying, and automatically generating counterspeech, to reduce the huge burden of manually producing it. In particular, researchers have taken different directions in addressing these challenges, thus providing a variety of related tasks and resources. In this paper, we provide a guide for doing research on counterspeech, by describing—with detailed examples—the steps to undertake, and providing best practices that can be learnt from the NLP studies on this topic. Finally, we discuss open challenges and future directions of counterspeech research in NLP.

**Content warning: this paper contains unobfuscated examples some readers may find offensive**

## 1 Introduction

Online spaces provide fertile ground for the diffusion of hateful content, which is often interlinked with episodes of offline violence (Awan and Zempi, 2016). Both witnessing and receiving hateful content can be detrimental to the mental health of victims and create a sense of insecurity (Saha et al., 2019; Siegel, 2020; Dreißigacker et al., 2024), determining the need to mitigate hate. In this context, counterspeech represents a promising strategy to oppose online hate, since it can be more effective than other moderation procedures (Benesch, 2014; Schieb and Preuss, 2016), while also protecting free speech (Kiritchenko et al., 2021). Because of its potential effectiveness, counterspeech has been investigated by non-governmental organisations


(NGOs) as a possible strategy to fight online hate. An example of hate speech (HS) and counterspeech (CS) from Fanton et al. (2021) is shown here:<sup>1</sup>

HS: Women are basically childlike, they remain this way most of their lives. Soft and emotional. It has devastated our once great patriarchal civilizations.

CS: Without softness and emotions there would be just brutality and cruelty. Not all women are soft and emotional and many men have these characteristics. To perpetuate these socially constructed gender profiles maintains patriarchal norms which oppress both men and women.

Given the amount of hateful content produced, an increasing number of Natural Language Processing (NLP) studies have begun to address the task of automatic counterspeech classification and generation. However, the settled definitions and best practices required to unify these efforts are still missing. While prior surveys largely focused on the effectiveness of deploying counterspeech in the real world (Chaudhary et al., 2021; Adak et al., 2022; Alsagheer et al., 2022; Chung et al., 2023), we offer a complete step-by-step guide on how to conduct NLP research on counterspeech for both newcomers and experts. In particular, we extensively review existing NLP studies and resources on counterspeech, propose common concepts and best practices, and point out the limitations and open challenges of what has been done so far. After providing some background (§2), the guide is articulated in three steps: task design, data selection and evaluation (§3, §4, §5, respectively).<sup>2</sup> Finally, we discuss the open challenges in the field. A complete description of the review methodology we used is provided in Appendix A.1.

<sup>1</sup>Throughout the paper, examples are coded with the following colour boxes: red for hate speech, light blue for counterspeech, and grey for everything else.

<sup>2</sup>These sections contain practical recommendations and best practices, marked with the spanner symbol .

## 2 Background

To better frame the concept of counterspeech we review definitions that have been proposed for it, the strategies that can be adopted, and several related tasks.

### 2.1 Definitions

The most common definition of counterspeech is that of [Benesch \(2014\)](#) and [Schieb and Preuss \(2016\)](#), who identify it as non-aggressive textual feedback that uses credible evidence, factual arguments and alternative viewpoints. Other works have focused on the relational nature of counterspeech: it only exists in response to hate speech ([Mathew et al., 2019](#); [Ashida and Komachi, 2022](#)), challenging, condemning it or providing an alternative viewpoint ([Vidgen et al., 2021](#); [Hangartner et al., 2021](#)). Also, it should explicitly condemn hate or support an abused entity ([He et al., 2021](#); [Vidgen et al., 2020](#)). Finally, it is consequences-oriented: it should discourage hate speech ([Rieger et al., 2018](#)) and aim to change what people think ([Qian et al., 2019](#)).

Although the terms *counterspeech* and *counter narrative* both rely on the idea that “the strategic response to hate speech is more speech” ([Bielefeldt et al., 2011](#)), in the social sciences *counter narratives* are representations that challenge dominant views in the areas of education, propaganda and public information ([Benesch et al., 2016b](#)).<sup>3</sup> Nevertheless, in the NLP studies included in this survey, these terms have been used interchangeably. Accordingly, we analyse works focusing on both “counter narratives” and “counterspeech”, but use the latter term, which we consider to be more appropriate.

### 2.2 Strategy taxonomies

Counterspeech can be distinguished by the strategy (or strategies) it employs. The most common taxonomy is that proposed by [Benesch et al. \(2016b\)](#), who distinguish seven types of counterspeech: *Presenting facts to correct misstatements or misperceptions*, *Pointing out hypocrisy or contradictions*, *Warning of consequences*, *affiliation*, *Denouncing hateful speech*, *Humor and sarcasm* and *Tone*. [Mathew et al. \(2019\)](#) split the latter category into *Positive* and *Hostile language*, and [Chung et al. \(2019\)](#) add *Counter-questions* on top of these.

<sup>3</sup>We refer to [Chung et al. \(2023\)](#) for a more detailed analysis of this distinction.

Other taxonomies have been proposed by [Qian et al. \(2019\)](#) and [Vidgen et al. \(2020\)](#): see Appendix A.2 for more details. However, not all strategies are equally effective: using *Hostile tone* can backfire, or discourage other counterspeakers from joining a conversation ([Benesch et al., 2016a](#)). [Mathew et al. \(2019\)](#) show how this type of counterspeech is not well-accepted even by the communities in whose favour it is produced, and provide this example:

CS: This is ridiculous!!!!!! I hate racist people!!!! Those police are a\*\*holes!!!

Similarly, [Benesch et al. \(2016a\)](#) advise that *Warning of consequences* should never turn into threats, as they show in this positive example:

CS: Current and future employers will be able to see your tweets, using the hashtag created to attack the chancellor of your university, with misogynist and racist content.

An empathetic, polite and constructive tone is also encouraged in guidelines written by counterspeech movements such as *Get the Trolls out*.<sup>4</sup>

In subsection 3.1, we discuss the task of automated identification of the strategies discussed here.

### 2.3 Related tasks

To better define counterspeech we describe its similarities and differences to several related tasks.<sup>5</sup> The first of these is **hope speech**, which indicates comments with a constructive view of the future and a peace-seeking intent ([Palakodety et al., 2019](#); [Chakravarthi, 2020](#); [Kumaresan et al., 2023](#); [García-Baena et al., 2023](#); [Jiménez-Zafra et al., 2023](#)). However, as opposed to counterspeech, hope speech does not necessarily reply to hate speech or challenge a message with an opposing stance:

Context: Video about COVID-19.

Reply: We will survive these things.

An overlapping phenomenon with hate speech is online trolling: an aggressive online behaviour seeking strong reactions for the sake of amusement ([Lee et al., 2022a](#)). Various **counter-trolling** strategies have been proposed. However, one of these consists in *reciprocating*, while counterspeech writing guidelines discourage direct attacks on the interlocutor:

<sup>4</sup>“Stopping hate: how to counter hate speech on Twitter”.

<sup>5</sup>A non-exhaustive list of available datasets for these tasks can be found in Appendix A.3.

Context: You gotta be a fucking idiot to be a sports fan lmao

Reply: Aren't u that guy who has to give ur neighbors the biannual update that ur a sex offender

Fraser et al. (2021) instead, address stereotypes by generating **anti-stereotype** words rather than true counterspeech while Allaway et al. (2022) provide counterstatements to essentialist beliefs (i.e. generalizations about a group):

Context: I speak English, I don't speak libt\*rd

Reply: Conservatives can also be stupid.

Even if counterspeech can contain anti-stereotyping messages, not all counterspeech strategies, such as asking *clarifying questions* (e.g. *What do you mean by "libt\*rd"?*), address the essentialist belief implicit in a text.

Other existing work has focused on making models better while handling dangerous or unsafe situations. Kim et al. (2022a) present a dataset of **prosocial dialogues**, i.e. conversations in which the speaker responds to potentially unsafe situations by promoting respect for social norms. As opposed to counterspeech, these responses are directed to unsafe rather than hateful texts, as in this example:

Context: I saw someone overdose and didn't tell anyone.

Reply: First of all, that must have been a terrifying situation. It was your responsibility to help the person in need though. If people find out that you saw the person overdose and didn't help, they will certainly not think highly of your behaviour either.

Alternatively, **detoxification** (Laugier et al., 2021; Logacheva et al., 2022) consists in rephrasing toxic comments with civil wording: however, this does not address any possible underlying hateful content, as shown by the following example.

Context: you now have to defend this clown along with his jewish corruption.

Detoxified: you now have to defend this guy from his jewish ties.

Ung et al. (2022) make models open in receiving **feedback** from users **about safety failures** of their generated content. Although this feedback resembles *denouncing* counterspeech, they tackle model-generated rather than user-generated hate speech:

Context: I am getting a kick out of watching you try to think you have value in the family.

Reply: no need to attack someone because you think differently.

Another related task is **counter-argumentation** generation (Hua and Wang, 2018, 2019; Hua et al., 2019; Alshomary et al., 2021; Alshomary and Wachsmuth, 2023). Still, a logically valid counter-argument is not necessarily a good counterspeech, as shown in this example from Fanton et al. (2021):

Context: We should kill all the jews.

Reply: There are many alternatives to removing jews, such as converting them to another religion (e.g. Buddhism).

Finally, **misinformation countering** consists of justifying the veracity of a statement (Stambach and Ash, 2020; Kotonya and Toni, 2020; Jolly et al., 2022; Ma et al., 2023; He et al., 2023a; Russo et al., 2023a,b).<sup>6</sup> These justifications can have some characteristics in common with counterspeech, e.g. being polite, fluent and relevant (He et al., 2023a; Russo et al., 2023a). However, counterspeech does not always contain evidence, and a factually inaccurate claim is not necessarily hateful, as shown in this example from Russo et al. (2023a):

Context: 11,000 of 13,000 knife attacks in London were carried out by Muslim migrants.

Reply: This claim is baseless as information on offenders' religion and nationality is not held by the authorities. Regardless, the claim is implausible.

### 3 Step 1: Design your task

The first step is to select which counterspeech task(s) to tackle. We discuss studies covering classification, selection and generation, and derive possible best practices from them.

#### 3.1 Classifying counterspeech

Classification can help to understand counterspeech dynamics and to collect counterspeech data. We consider three sub-tasks.

**CS detection.** Several works focus on detecting counterspeech as opposed to: non-counterspeech (Mathew et al., 2019; Goffredo et al., 2022; Al-banyan et al., 2023a), hate speech (Garland et al., 2020), hate speech and neutral instances (Möhle


<sup>6</sup>We refer readers to He et al. (2023b)'s survey, which analyses approaches to crowd-based and effective counter-misinformation.

et al., 2023; Yu et al., 2022; Shah et al., 2022; He et al., 2021; Vidgen et al., 2021), and among Hostility, Criticism and Non-related instances (Vidgen et al., 2020). Finally, Goffredo et al. (2022) also identified messages supporting counterspeech.

**User classification.** Only Mathew et al. (2020) worked on classifying Twitter users into hateful or counterspeakers: this task can be useful for a platform to intervene early and demote hateful accounts, while promoting counterspeech.

**Strategy classification.** Detecting the counterspeech strategies<sup>7</sup> used (Mathew et al., 2019; Chung et al., 2021a) can help to analyse their effectiveness and develop more fine-grained responses. Similarly, Albanyan and Blanco (2022) identify counterspeech, determining whether it provided a justification, attacks the author of the hate speech, or includes additional hate.


While some of these classification studies employ only traditional classifiers (Mathew et al., 2020; Shah et al., 2022), others compare them with neural models, showing that the latter perform better or comparably well than the first (Mathew et al., 2019; He et al., 2021; Vidgen et al., 2021). Most studies employ only neural models and experiment with different types of input, showing how including the context (e.g. the hate speech) helps to reduce false negatives (Vidgen et al., 2021; Yu et al., 2022; Albanyan and Blanco, 2022). In fact, hate speech and counterspeech can share similar textual features to some extent, making it difficult to automatically distinguish them without further context (Möhle et al., 2023). Better counterspeech detection performance is obtained by pretraining the models on similar tasks, such as stance (Yu et al., 2022) or emotion detection (Albanyan and Blanco, 2022).

 The most common errors in counterspeech classification arise when the text is complex and contains irony or sarcasm (Goffredo et al., 2022; Albanyan and Blanco, 2022), negation (Yu et al., 2022), or when more context is needed to disentangle counterspeech from other categories (Vidgen et al., 2021). Another problem, common to hate speech detection, is lexical overfitting to specific terms or swearwords (Vidgen et al., 2020; Yu et al., 2022). In other cases, errors might arise from the annotation itself (Vidgen et al., 2020). Using a large enough dataset with high-quality annotation can help to reduce such errors.

<sup>7</sup>See subsection 2.2 for an overview of counterspeech strategy taxonomies.

### 3.2 Selecting counterspeech responses

One way to produce counterspeech consists of selecting from a pool of possible responses that can be obtained via over-generation (Zhu and Bhat, 2021). Alternatively, the candidates can be retrieved from a counterspeech dataset: Chung et al. (2021c) rely on a *tf-idf* information retrieval model, while Akazawa et al. (2023) employ the implicit stereotype of the hate speech to make a selection via cosine similarity. It is also possible to select counterspeech among non-counterspeech content available online, e.g. from Twitter (Möhle et al., 2023) or online articles (Albanyan et al., 2023a).

 Filtering a social media dataset containing both counterspeech and non-counterspeech instances does not produce a larger amount of counterspeech than a random sample (Möhle et al., 2023). Thus, selection seems particularly useful to obtain the most appropriate response to a specific hate speech when a pool of gold (Akazawa et al., 2023; Chung et al., 2021c) or silver (Zhu and Bhat, 2021) counterspeech is already available, rather than filtering out non-counterspeech instances.

### 3.3 Generating counterspeech

Suitable counterspeech can take many forms: we outline non-exhaustive desirable aspects of counterspeech (knowledge, personality, style), and report relevant techniques for generation (fine-tuning and prompting, translation).

**Knowledge guided generation.** Both Chung et al. (2021b) and Jiang et al. (2023) structure this task in two phases: first the extraction of relevant knowledge from an external source, and secondly the generation of knowledge-augmented counterspeech. For the first phase, Chung et al. (2021b) used extracted keyphrases to select sentences from Wikipedia articles and news datasets, while Jiang et al. (2023) rely on stance consistency, semantic overlap rate, and fitness for hate speech to construct a knowledge repository from the ChangeMyView subreddit.

**Personality guided generation.** Examples of this approach are de los Riscos and D’Haro (2021), who employed the PersonaChat dataset to fine-tune a model provided with a dynamic persona profile or dialogue history as input during generation, and Doğanç and Markov (2023), who experimented with both fine-tuning and few-shot prompting to incorporate the profiling information and obtain personalized counterspeech.

**Style guided generation.** Here, we include all other stylistic features addressed during generation. To enhance *specificity*, Bonaldi et al. (2023) employ two attention-based regularization techniques to include a broader context during training and generation, while Furman et al. (2023a) focus on the *argumentative information* present in the hate speech to guide the generation towards particular response strategies. Other works target multiple aspects at the same time: Saha et al. (2022) simultaneously control for the *politeness*, *detoxification* and *emotion* in the generated counterspeech. Finally, Gupta et al. (2023) propose a two stage-framework for generating counterspeech conditioned on five different *strategies* (i.e. informative, denouncing, question, positive, and humour).

Despite the importance of knowledge-driven generation, correcting misinformation alone is not sufficient and can lead to higher levels of violence (Carthy and Sarma, 2023). For this reason, taking into consideration other aspects is fundamental: for example, Hangartner et al. (2021) showed how empathy-based counterspeech can have an impact, however small, in reducing hate speech. Moreover, generating counterspeech with specific strategies according to the targeted community can be particularly effective, and in general, maintaining a polite tone is recommended (Mathew et al., 2019).

**Fine-tuning and prompting.** The most commonly employed approach for counterspeech generation is fine-tuning a language model on a counterspeech dataset (e.g. Qian et al., 2019; Tekiroglu et al., 2022; Halim et al., 2023). However, recent advances have allowed generation of counterspeech via few-shot (Ashida and Komachi, 2022; Furman et al., 2023a; Vallecillo-Rodríguez et al., 2023; Doğanç and Markov, 2023), one- and zero-shot prompting (Mun et al., 2023; Zheng et al., 2023).

**Translation and low-resourced languages.** Chung et al. (2020) generate Italian counterspeech by fine-tuning a model on a combination of gold and silver Italian data obtained via translation. Also Vallecillo-Rodríguez et al. (2023) rely on translated examples from Chung et al. (2021b) to create a Spanish corpus via few-shot prompting. Finally, Furman et al. (2023a) include Spanish examples in their generation task.

Prompting allows generation of counterspeech in a low computationally intensive way: however, given the specificity of the task, few-shot prompting is preferred over one- and zero-shot prompting. Moreover, clear and specific instructions should be given to the model to obtain more fine-grained replies. In particular, both Hassan and Alikhani (2023) and Mun et al. (2023) show how LLMs tend to use general strategies such as *denouncing*, *comment* or *correction* when generating counterspeech without specific indications. Another viable strategy to obtain data in low-resourced scenarios is translation, as shown by Chung et al. (2020) and Vallecillo-Rodríguez et al. (2023), who respectively use silver translated data alone or together with gold data in the language of interest to generate responses in Spanish and Italian.

## 4 Step 2: Select the data

After task design, the next choice is whether to collect a new dataset or to use an already existing one. We will discuss the use-cases of the main counterspeech collection procedures, and then detail the characteristics of available counterspeech datasets.

### 4.1 Collecting your own data

Collecting data entails specific consequences according to the chosen strategy (Tekiroğlu et al., 2020): we summarise them below and in Table 1.

Coll.	Data type	Quant.	Conf.	Div.	Non-eph.
Crawl.	Real	✓	-	✓	-
Crowd.	Simulated	✓	✓	-	✓
Niche.	Simulated	-	✓	✓	✓
Hybr.	Synthetic	✓	✓	✓	✓
Auto.	Synthetic	✓	✓	-	✓

Table 1: Data type, quantity, conformity to counterspeech writing guidelines, diversity and non-ephemerality of counterspeech collected with different procedures.

**Crawling.** This consists of scraping real counterspeech from sources such as Youtube (Mathew et al., 2019), Twitter (e.g. Mathew et al., 2020; Vidgen et al., 2020; Goffredo et al., 2022), Reddit (Yu et al., 2022; Vidgen et al., 2021; Hassan and Alikhani, 2023), and online articles (Albanyan et al., 2023a). Crawling can guarantee high quantity and diversity of data. However, responses in the wild do not always conform to counterspeech guidelines, as shown in this example from Albanyan and Blanco (2022):

HS: on my way to f\*\*k your b\*\*ch.

CS: But my b\*\*ch is your mamma

To tackle such problems, Hassan and Alikhani (2023) manually remove offensiveness and first-person references from their crawled data. Furthermore, relying on lists of keywords or users may not capture the entirety of available counterspeech. Finally, crawled data might be ephemeral, hindering reproducibility<sup>8</sup> (Klubicka and Fernández, 2018).

**Crowdsourcing.** Qian et al. (2019) and Furman et al. (2022, 2023a) rely on non-expert annotators to manually produce counterspeech. With crowdsourcing (Crowd.) it is possible to obtain a high quantity of data, although these often contain simple and poorly varied arguments. For example, a large amount of human-written counterspeech can consist in *denouncing* the hate speech: even if this is an acceptable strategy, it represents a simple form of reply, as shown in this example from Qian et al. (2019):

HS: Like... I literally addressed this point and you completely ignore it, like a stupid retard nigger.

CS: Stop using racist and medical conditions as an insult.

**Nichesourcing.** This relies on experts to produce counterspeech (Chung et al., 2019, 2021b). Nichesourcing (Niche.) can guarantee the highest quality data in terms of syntactic variety and semantic complexity. However, it is also the most expensive strategy, making it difficult to obtain a large quantity of data.

**Hybrid and fully automated collection.** Hybrid approaches (Hybr.) exploit a combination of automatic and manual intervention to collect data. Tekiroğlu et al. (2020) propose an author-reviewer architecture where a generation model (author) produces hate speech and counterspeech data, and human reviewers filter and post-edit the generated data. This strategy is then employed by both Fanton et al. (2021) and Bonaldi et al. (2022a). Below, an example of counterspeech post-editing from Fanton et al. (2021) is shown:

HS: Transgenders want to rape our children.

CS: This is not true. Maybe they are worried because of the rise in hate crimes, incidents of which are down to 28 percent, since 2014.

CS<sub>pe</sub>: This is not true. Maybe *you should be worried* about the rise in hate crimes *against queers*, incidents of which *are almost doubled* since 2014.

<sup>8</sup>Albanyan and Blanco (2022) and Albanyan et al. (2023b) include only the Tweets IDs and not the full texts.

Ashida and Komachi (2022) and Vallecillo-Rodríguez et al. (2023) fully rely on generative models to produce counterspeech (Auto.): human annotators do not intervene.

🔍 *Crawling* is the most common data collection procedure used in the wild to gather counterspeech. However, *nichesourcing* can generate the highest-quality responses, since it benefits from expert knowledge. If expertise is limited, partially automatising data collection via a *combination* of a fine-tuned model and human post-editing can be a good solution. If expertise is extremely limited, non-expert annotators or a classifier (Hassan and Alikhani, 2023) can prefilter the data prior to expert validation (Tekiroğlu et al., 2020). Alternatively, non-expert annotators can be trained, following the procedure described in Appendix A.4. However, we discourage relying solely on automatic counterspeech collection without human intervention, given the sensitivity of this task.

## 4.2 Choosing from existing datasets

An efficient alternative to data collection is selecting among available counterspeech datasets. We describe them along several dimensions, summarised in Table 2, to facilitate the choice of most suitable dataset for specific research needs.

**Shape of the interactions.** Available datasets can be divided into four main groups according to the type of interaction they contain. Single comments (Single c.) are individually labeled as hate speech, counterspeech, or other classes, without further conversational context, and often come from social media platforms such as Twitter or Reddit (Vidgen et al., 2020, 2021; He et al., 2021). Pairs of hate speech and their related counterspeech are the most widely diffused type of interaction encoded in available datasets (e.g. Chung et al., 2019; Goffredo et al., 2022; Vallecillo-Rodríguez et al., 2023). Alternatively, pairs with context (Pairs+c.) include a longer conversational context, such as previous or subsequent comments (Mathew et al., 2019; Qian et al., 2019; Albanyan et al., 2023b). Finally, Bonaldi et al. (2022a) present hate speech and counterspeech dialogues (Dialog.) including multiple counterspeech turns.

**Targets of hate.** Most studies include multiple targeted minorities: the most represented are *Jews*, *Blacks*, and *LGBT* (Mathew et al., 2019). Additionally, Chung et al. (2021b), Bonaldi et al. (2022a) and Vallecillo-Rodríguez et al. (2023) consider *Migrants*, *Muslims*, and *Women*, Hassan and Alikhani (2023) cover *Disabled* people, and Fanton et al. (2021) include *Overweight* and *Romani* people on

Dataset	Size	# CS	Interact.	Coll.	Source	Lang.	Tar.	Add.
Mathew et al. (2019)	13,924	6,898	Pairs + c.	Crawl.	YouTube	EN	✓	✓
Chung et al. (2019)	14,988	14,988	Pairs	Nich.	NGOs op.	EN/FR/IT	✓	✓
Qian et al. (2019)	16,845	29,388	Pairs + c.	Crowd.	Reddit, Gab	EN	-	-
Mathew et al. (2020)	1,290	1,290	Pairs	Crawl.	Twitter	EN	-	✓
Vidgen et al. (2020)	20,000	116	Single c.	Crawl.	Twitter	EN	✓	-
He et al. (2021)	2,290	517	Single c.	Crawl.	Twitter	EN	✓	✓
Vidgen et al. (2021)	27,494	220	Single c.	Crawl.	Reddit	EN	-	-
Chung et al. (2021b)	195	195	Pairs	Niches.	NGO op.	EN	✓	✓
Fanton et al. (2021)	5,003	5,003	Pairs	Hybr.	NGOs op.	EN	✓	-
Yu et al. (2022)	6,846	1,622	Pairs	Crawl.	Reddit	EN	-	✓
Albanyan and Blanco (2022)	5,652	1,149	Pairs	Crawl.	Twitter	EN	-	✓
Bonaldi et al. (2022a)	3,059	8,311	Dialog.	Hybr.	NGOs op.	EN	✓	-
Ashida and Komachi (2022)	348	306	Pairs	Autom.	Autom.	EN	-	✓
Goffredo et al. (2022)	624	81	Pairs	Crawl.	Twitter	IT	✓	✓
Furman et al. (2022)	2,055	2,055	Pairs	Crowd.	Basile et al. (2019)	ES	-	✓
Furman et al. (2023a)	2,077	2,077	Pairs	Crowd.	Furman et al. (2023b)	EN/ES	-	-
Vallecillo-Rodríguez et al. (2023)	238	238	Pairs	Autom.	Chung et al. (2021b)	ES	✓	✓
Hassan and Alikhani (2023)	3,900	250	Pairs	Crawl.	Reddit	EN	✓	✓
Albanyan et al. (2023b)	2,621	1,685	Pairs + c.	Crawl.	Twitter	EN	-	✓
Albanyan et al. (2023a)	54,816	2,365	Pairs	Crawl.	Web articles	EN	✓	-

Table 2: Available datasets, according to their size, nr. of counterspeech interaction type, data collection procedure, source, language, target, and additional information. The data size and the number of counterspeech refer to the interactions shape (e.g. 5,003 *pairs*), except for Qian et al. (2019) and Bonaldi et al. (2022b) where the number of effective counterspeech turns is shown.

top of these. Other studies focus on a single target: in particular, Chung et al. (2019) on *Islamophobia*, whereas He et al. (2021) and Vidgen et al. (2020) on COVID-19 related *Asian* hate. Only Fanton et al. (2021) include, in a few examples, intersectional hate, i.e. hate directed towards people belonging to multiple minorities, such as *black women*. Finally, Albanyan et al. (2023a) is the only research to address hate towards *individuals*, rather than groups.

**Types of hate addressed.** Chung et al. (2019) identify hate speech according to the sub-topic it covers: *culture, economics, crimes, rapism, women oppression, history* and *other/generic*. Vidgen et al. (2020) make different levels of distinctions according to the offensiveness intensity (*Hostility* and *Criticism*) and category (*Interpersonal abuse, Use of threatening language* or *Dehumanization*). Finally, Vidgen et al. (2021) distinguish hate according to the addressed entity (an *identity*, an *affiliation* or an *identifiable person*) and category (*derogation, animosity, threatening, dehumanization* or *glorification of hateful entities*).

**Languages.** Most existing datasets are in English, with only a few covering French (Chung et al., 2019), Italian (Chung et al., 2019; Goffredo et al., 2022), and Spanish (Furman et al., 2022, 2023a;

Vallecillo-Rodríguez et al., 2023).

**Additional information.** Other information present in these datasets include the counterspeech strategy (Mathew et al., 2019; Chung et al., 2019; Mathew et al., 2020; Goffredo et al., 2022, see Section 2.2). Similarly, Albanyan and Blanco (2022) specify whether the counterspeech contains a justification or attacks the author, and if it is not a counterspeech whether it agrees with the hater or adds additional hate. Albanyan et al. (2023b) do the same but for replies to counterspeech. Others have considered the *discourse*<sup>9</sup> (Hassan and Alikhani, 2023) and the *argumentative strategy* countering the hate speech (Furman et al., 2022, 2023a).

Then, there can be contextual information on social media platforms data, such as the title of the discussion (Yu et al., 2022), the list of replies and the timestamp (Mathew et al., 2019), the number of likes and replies (Mathew et al., 2019), or the ego network of the users (He et al., 2021).

Other aspects are the annotator demographics, (Chung et al., 2019), or the human annotations: Ashida and Komachi (2022) and Vallecillo-Rodríguez et al. (2023) include the counterspeech offensiveness, stance and informativeness. Further

<sup>9</sup>From the Segmented Discourse Representation Theory (Asher and Lascarides, 2003)

fine-grained information include knowledge sentences (Chung et al., 2021b) and paraphrases of crawled counterspeech via the removal of offensiveness and first-person references (Hassan and Alikhani, 2023).

✎ The choice of the dataset should be driven by task design. It is important to consider the dataset size and the actual number of counterspeech instances: a few examples can be enough for few-shot prompting, while larger datasets are beneficial for fine-tuning or selection tasks. Additionally, the source and procedure of data collection can affect the structure, style, and strategies of the included counterspeech (e.g. Tweets are shorter, crowdsourced data often contain *denouncing* counterspeech). Finally, any additional information can be decisive according to the specific goal of the study, e.g. the knowledge sentences in the dataset by Chung et al. (2019) for knowledge-driven generation.

## 5 Step 3: Evaluate

Next, we look at the literature on evaluating counterspeech based on the tasks discussed in §3.<sup>10</sup>

### 5.1 Evaluating classification

When gold test data is available, performance can be assessed via F1, precision, recall, accuracy, and confusion matrices. Moreover, in multi-label scenarios (e.g. counterspeech employing multiple strategies), hamming loss is recommended as it can better capture model performance by considering the ratio of true classes in a prediction rather than a hard right prediction (Mathew et al., 2019). Finally, human judgement can be compared to classifiers to verify their performance (Garland et al., 2020), and qualitative error analysis can help to better understand the specific flaws of a model (Vidgen et al., 2020, 2021; Goffredo et al., 2022; Yu et al., 2022).

### 5.2 Evaluating generation

Standard evaluation metrics can be grouped into extrinsic (measuring the potential impact of a system on its related tasks or on achieving its overall goals) and intrinsic measures (assessing the system output in isolation, Walter, 1998).

**Extrinsic evaluation.** So far, only Chung et al. (2021c) have focused on this kind of evaluation. To assess how effective their counterspeech suggestion tool was in empowering NGO operators during hate countering, the operators were asked to evaluate their user experience through a questionnaire

<sup>10</sup>The metrics described in §5.2 are meant for evaluating generation tasks, but they can be used for selection too.

(Laugwitz et al., 2008) and open-ended qualitative questions.<sup>11</sup>

**Intrinsic automatic metrics.** Some of these metrics centre on the comparison between generation and references using criteria such as linguistic surface (Papineni et al., 2002; Lin, 2004), novelty (Wang and Wan, 2018), and semantic similarity (Zhang et al., 2019). Furthermore, some work measures the quality of counterspeech generation based on fine-grained characteristics, such as toxicity (Google Jigsaw, 2022), informativeness (Fu et al., 2023), factuality (Fu et al., 2023), repetitiveness (Bertoldi et al., 2013; Cettolo et al., 2014), linguistic acceptability, politeness, emotion (Saha et al., 2022), stance and relevance to the input (i.e. the hate speech, Schütze, 2008; Halim et al., 2023).

**Human evaluation.** Several factors should be considered, such as evaluation criteria, scale (e.g. ranking vs. Likert or sliding scale), and annotators (e.g. experts vs. crowd). The common approach is to ask annotators to judge responses on a scale (e.g. of 1 to 5) based on aspects including suitability and specificity (Chung et al., 2021b; Tekiroğlu et al., 2022; Bonaldi et al., 2023), grammaticality (Chung et al., 2020; Zhu and Bhat, 2021), coherence and informativeness (Chung et al., 2021b).

✎ While intrinsic automatic metrics can capture the overall performance of generation systems at scale, some of these lack interpretability and correlation with human evaluation (Belz and Reiter, 2006; Novikova et al., 2017). Considering the complexity of hate mitigation, human evaluation is a more reliable approach. Most previous work uses experts or trained annotators for manual evaluation. Since the choice of the best response is subjective, it is desirable to enlist diverse annotators (e.g. in regard to gender and educational level, Waseem et al., 2017; Sap et al., 2019; Abercrombie et al., 2023) or users identifying with the potential recipients of counterspeech such as perpetrators and bystanders.

## 6 Open challenges

Drawing from the surveyed literature, we highlight key open challenges in counterspeech research.

**Language and culture.** Hate speech is not only linguistically, but also culturally specific. Therefore, it requires culturally specific responses. For example, in Spanish, the same words can convey discriminatory connotations depending on the country in which they are used (Castillo-López et al.,

<sup>11</sup>For a detailed discussion on evaluating the impact of counterspeech in real-life scenarios, see Chung et al. (2023).



2023). Moreover, the same groups can be subject to different stereotypes associated with the historical events of their location (Laurent, 2020).

**Sources of hate.** A level of granularity not yet considered for counterspeech design is the identity of the hate speech perpetrator. This, in turn, can be considered together with cultural and geographical factors, to produce counterspeech tailored to specific targets (e.g. Italian neonazis).

**Types of hate.** Studies on counterspeech are mostly centred on explicit hate with only a few addressing stereotypes, prejudice or biases (Mun et al., 2023). Such implicit hate often contains complex linguistic forms with indirect sarcasm or humour (Waseem and Hovy, 2016; Fortuna and Nunes, 2018), and can be generic (“*boys play with trucks*”, Rhodes et al., 2012; Leslie, 2014), posing challenges in how to mitigate it (Buerger, 2022).

**Hallucinations.** Even if counterspeech does not necessarily need to contain factual evidence (§2.3), it can be effective in highlighting the absurdity of hate speech. However, a challenge in open-ended generation is hallucinations.<sup>12</sup> One way to address this is to rely on external knowledge sources (Chung et al., 2021b; Jiang et al., 2023): here, RAG systems (Lewis et al., 2020; Ram et al., 2023) are a promising research direction. Alternatively, inaccurate text can be detected in the generation (Manakul et al., 2023). Finally, counterspeech should be placed in the right temporal context to be more effective: knowledge-grounded generation can help to produce more time-relevant responses.

**Evaluation.** As discussed in Section 5, existing evaluation metrics are limited. It would be desirable to create test suites analysing different functionalities of counterspeech generation models; e.g. testing models’ capacity to generate counterspeech directed at specific types of hate with certain strategies (similar to the HateCheck initiative, Röttger et al., 2021). Additionally, the definition of good counterspeech is subjective and should be user-oriented (e.g. assessed by the target audience). Hence, an ideal evaluation could involve gathering multiple perspectives on suitable counterspeech.

**Biases in data collection.** Possible biases can emerge from various choices taken during data collection. Firstly, the data source can strongly affect

content and style. With crawling, collecting texts from a specific platform will determine its length, style and topics, mainly representing the users of that platform and thus not being highly generalisable. With crowdsourcing and nichesourcing, it should be considered that annotators have different sensitivity to hate, according to their country of origin (Lee et al., 2023), their belonging to a targeted minority, and their personal experiences. This can have a considerable impact on the content of the counterspeech they write too. Moreover, non-experts recur more often to simpler counterspeech strategies such as *denouncing* than experts (Tekiroğlu et al., 2020). Choice of annotators also creates similar possible biases to human evaluation, as discussed above and in section 5.2: for all these reasons, it is better to recruit a diverse set of annotators, if possible. Finally, bias can also originate from the other factors already discussed, i.e. the considered targets of hate, language and geographical/temporal context of the collected data. In general, it is always preferable to provide newly introduced datasets with a dataset card<sup>13</sup> to inform users on how to responsibly employ the data, limiting the emergence of possible harms.

## 7 Conclusion

We presented a thorough review of 43 NLP studies on counterspeech. This is organised as a step-by-step guide, intended for those approaching counterspeech from an NLP perspective. First, we framed counterspeech and its strategies, distinguishing it from other similar tasks. Then, we structured the subsequent sections as progressive steps to undertake when approaching counterspeech research in NLP: in these sections, we relied on the literature to provide insights into the consequences each choice might imply. Finally, we point out open challenges in the field. Counterspeech represents a promising approach to tackling online hate, and NLP can potentially provide the tools to make it scalable. However, an efficient system is not necessarily a good system: researchers operating in this area must be aware of the consequences entailed by each of their choices, to avoid spreading further harm.

<sup>12</sup>I.e. text nonsensical or unfaithful to the provided source input (Ji et al., 2023), often with factually incorrect content.

<sup>13</sup><https://huggingface.co/docs/hub/datasets-cards>

## Limitations

To an external reader, the number of papers included in this study might seem small: this is both because relatively little attention has still been devoted to this topic, and because we made the specific choice of focusing only on NLP papers proposing one or more of the following contributions: a dataset, a classification, selection or generation task. In this survey, we included studies from Scopus, arXiv and the ACL Anthology, following the methodology of previous abusive language surveys in the NLP domain (Chung et al., 2023; Vidgen and Derczynski, 2020). Our search was conducted using keywords, which might not be comprehensive of all the available studies on counterspeech but represents a reasonable compromise for searching in such huge databases. Moreover, all the authors already had research experience in counterspeech and thus had a personal list of counterspeech studies collected over the years—all of which were retrieved with the automated search process.

## Ethical considerations

In addition to potential legal issues (see Chung et al., 2023, for a discussion of this), engaging in counterspeech has important social consequences: for this reason, many precautions should be adopted when dealing with it, similarly to other abusive language related domains. First of all, researchers and potential annotators involved in any counterspeech task should prioritise their mental well-being: prolonged exposure to abusive content can have negative effects, that can be avoided by following the mitigation measures described by Vidgen et al. (2019a) and Kirk et al. (2022) (see Appendix A.4 for more details). For what regards data collection and distribution, synthetic data represent a viable option to preserve users privacy. Moreover, using simulated hate speech that are simple and stereotyped can avoid possible negative outcomes such as training a language model on hate speech generation. However, if the collected data are real, it is important to ensure that this does not interfere with the online activities of counterspeakers. For example, if the counterspeech included in a dataset are obtained by scraping from a list of activists' accounts, malicious users might reverse-search these texts, and identify the operators' accounts, thus exposing them to possible attacks. Finally, regarding the deployment of generation systems in real-life scenarios, human supervision is still necessary: the

risks of hallucinations and abusive generation are still too high to fully automate the task of counterspeech production in the wild.

## Acknowledgements

Gavin Abercrombie was supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1). Yi-Ling Chung was supported by the Ecosystem Leadership Award under the EPSRC Grant EPX03870X1 & The Alan Turing Institute.

## References

- Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. [Resources for automated identification of online gender-based violence: A systematic review](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.
- Sayantan Adak, Souvic Chakraborty, Paramita Das, Mithun Das, Abhisek Dash, Rima Hazra, Binny Mathew, Punyajoy Saha, Soumya Sarkar, and Animesh Mukherjee. 2022. Mining the online infosphere: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1453.
- Nami Akazawa, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. [Distilling implied bias from hate speech for counter narrative selection](#). In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 29–43, Prague, Czechia. Association for Computational Linguistics.
- Abdullah Albanyan and Eduardo Blanco. 2022. Pinpointing fine-grained relationships between hateful tweets and replies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10418–10426.
- Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023a. [Finding authentic counterhate arguments: A case study with public figures](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13862–13876, Singapore. Association for Computational Linguistics.
- Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023b. [Not all counterhate tweets elicit the same replies: A fine-grained analysis](#). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (\*SEM 2023)*, pages 71–88, Toronto, Canada. Association for Computational Linguistics.
- Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90.
- Emily Allaway, Nina Taneja, Sarah-Jane Leslie, and Maarten Sap. 2022. Towards countering essentialism through social bias reasoning. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Dana Alsagheer, Hadi Mansourifar, and Weidong Shi. 2022. [Counter hate speech in social media: A survey](#).
- Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827.
- Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967.
- N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.
- Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. *WOAH 2022*, page 11.
- Imran Awan and Irene Zempi. 2016. The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and violent behavior*, 27:1–8.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.
- Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320.
- Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: United States Holocaust Memorial Museum*.
- Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016a. Considerations for successful counterspeech. *Dangerous speech project*.
- Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016b. Counterspeech on twitter: A field study. *Dangerous Speech Project*. Available at: <https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/>.
- Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.
- Heiner Bielefeldt, Frank La Rue, and Githu Muigai. 2011. Ohchr expert workshops on the prohibition of incitement to national, racial or religious hatred. In *Expert workshop on the Americas*.
- Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. [Weigh your own words: Improving hate speech counter narrative generation via attention regularization](#). In *Proceedings of the*

- 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 13–28, Prague, Czechia. Association for Computational Linguistics.
- Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022a. [Human-machine collaboration approaches to build a dialogue dataset for hate speech countering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022b. [Human-machine collaboration approaches to build a dialogue dataset for hate speech countering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049.
- Catherine Buerger. 2022. [Why they do it: Counterspeech theories of change](#). Available at SSRN 4245211.
- Sarah L Carthy and Kiran M Sarma. 2023. [Countering terrorist narratives: Assessing the efficacy and mechanisms of change in counter-narrative strategies](#). *Terrorism and Political Violence*, 35(3):569–593.
- Galo Castillo-López, Arij Riabi, and Djamé Seddah. 2023. [Analyzing zero-shot transfer scenarios across spanish variants for hate speech detection](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13.
- Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. [The repetition rate of text as a predictor of the effectiveness of machine translation adaptation](#). In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179.
- Bharathi Raja Chakravarthi. 2020. [HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion](#). In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, pages 41–53.
- Mudit Chaudhary, Chandni Saxena, and Helen Meng. 2021. [Countering online hate speech: An nlp perspective](#). *arXiv preprint arXiv:2109.02941*.
- Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. [Understanding counterspeech for online harm mitigation](#). *arXiv preprint arXiv:2307.04761*.
- Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021a. [Multilingual counter narrative type classification](#). In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. [CONAN - Counter Narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.
- Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. [Italian counter narrative generation to fight online hate speech](#). In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it*.
- Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. [Towards knowledge-grounded counter narrative generation for hate speech](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.
- Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021c. [Empowering NGOs in countering online hate messages](#). *Online Social Networks and Media*, 24:100150.
- Agustín Manuel de los Riscos and Luis Fernando D’Haro. 2021. [Toxicbot: A conversational agent to fight online hate speech](#). *Conversational dialogue systems for the next decade*, pages 15–30.
- Mekselina Doğanç and Ilia Markov. 2023. [From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech](#). In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 1–12, Prague, Czechia. Association for Computational Linguistics.
- Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. 2024. [Online hate speech victimization: consequences for victims’ feelings of insecurity](#). *Crime Science*, 13(1):4.
- Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. [Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240.
- Paula Fortuna and Sérgio Nunes. 2018. [A survey on automatic detection of hate speech in text](#). *ACM Comput. Surv.*, 51(4).
- Kathleen C Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. [Understanding and countering stereotypes: A computational approach to the stereotype content model](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. [Gptscore: Evaluate as you desire](#). *arXiv preprint arXiv:2302.04166*.

- Damián Furman, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023a. [High-quality argumentative information in low resources approaches improve counter-narrative generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, Singapore. Association for Computational Linguistics.
- Damian A Furman, Pablo Torres, Jose A Rodriguez, Lautaro Martinez, Laura Alonso Alemany, Diego Letzen, and Maria Vanina Martinez. 2022. Parsimonious argument annotations for hate speech counter-narratives. *arXiv e-prints*, pages arXiv–2208.
- Damián Ariel Furman, Pablo Torres, José A. Rodriguez, Diego Letzen, Maria Vanina Martinez, and Laura Alonso Alemany. 2023b. Which argumentative aspects of hate speech in social media can be reliably identified? In *Proceedings of Fourth International Workshop on Designing Meaning Representations, co-located with IWCS 2023*.
- Daniel García-Baena, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, and Rafael Valencia-García. 2023. Hope speech detection in spanish: The lgbt case. *Language Resources and Evaluation*, pages 1–28.
- Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112.
- Pierpaolo Goffredo, Valerio Basile, Bianca Cepollaro, and Viviana Patti. 2022. [Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 57–66, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Google Jigsaw. 2022. [Perspective API](#). Accessed: 26 May 2023.
- Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. [Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.
- Sadaf MD Halim, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani Thuraisingham. 2023. Wokegpt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.
- Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment.
- Sabit Hassan and Malihe Alikhani. 2023. [Discgen: A framework for discourse-informed counterspeech generation](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 420–429, Nusa Dua, Bali. Association for Computational Linguistics.
- Bing He, Mustaque Ahamad, and Srijan Kumar. 2023a. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.
- Bing He, Yibo Hu, Yeon-Chang Lee, Soyoung Oh, Gaurav Verma, and Srijan Kumar. 2023b. A survey on the role of crowds in combating online misinformation: Annotators, evaluators, and creators. *arXiv preprint arXiv:2310.02095*.
- Bing He, Caleb Ziemis, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94.
- Xinyu Hua, Zhe Hu, and Lu Wang. 2019. [Argument generation with retrieval, planning, and realization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.
- Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230.
- Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 591–602.
- Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).

- Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.
- Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, Daniel García-Baena, José Antonio García-Díaz, Bharathi Raja Chakravarthi, Rafael Valencia-García, and Luis Alfonso Ureña-López. 2023. Overview of hope at iberlef 2023: Multilingual hope speech detection. *Procesamiento del Lenguaje Natural*, 71:371–381.
- Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating fluent fact checking explanations with unsupervised post-editing. *Information*, 13(10):500.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022a. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022b. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.
- Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Filip Klubicka and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*, page 16.
- Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.
- Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Subalalitha Cn, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Momchil Hardalov, Ivan Koychev, Preslav Nakov, et al. 2023. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 47–53.
- Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.
- Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 63–76. Springer.
- Mario Laurent. 2020. Project hatemeter: helping ngos and social science researchers to analyze and prevent anti-muslim hate speech on social media. *Procedia Computer Science*, 176:2143–2153.
- Huije Lee, Young Ju NA, Hoyun Song, Jisu Shin, and Jong Park. 2022a. Elf22: A context-based counter trolling dataset to combat internet trolls. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3530–3541, Marseille, France. European Language Resources Association.
- Huije Lee, Young Ju Na, Hoyun Song, Jisu Shin, and Jong Park. 2022b. ELF22: A context-based counter trolling dataset to combat Internet trolls. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3530–3541, Marseille, France. European Language Resources Association.
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023. Crehate: Cross-cultural re-annotation of english hate speech dataset. *arXiv preprint arXiv:2308.16705*.
- Sarah-Jane Leslie. 2014. Carving up the social world with generics. *Oxford studies in experimental philosophy*, 1.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

- Yingchen Ma, Bing He, Nathan Subrahmanian, and Srijan Kumar. 2023. Characterizing and predicting social correction on twitter. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 86–95.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. [Interaction dynamics between hate and counter users on twitter](#). In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, CoDS COMAD 2020*, page 116–124, New York, NY, USA. Association for Computing Machinery.
- Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhanian, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. [Thou shalt not hate: Countering online hate speech](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.
- David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. 2010. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *International journal of surgery*, 8(5):336–341.
- Pauline Möhle, Matthias Orlikowski, and Philipp Ciminiano. 2023. [Just collect, don’t filter: Noisy labels do not improve counterspeech collection for languages without annotated resources](#). In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 44–61, Prague, Czechia. Association for Computational Linguistics.
- Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2242. Association for Computational Linguistics.
- Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2019. Hope speech detection: A computational analysis of the voice of peace. *arXiv preprint arXiv:1909.12940*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. [A benchmark dataset for learning to intervene in online hate speech](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.
- Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek. 2012. [Cultural transmission of social essentialism](#). *Proceedings of the National Academy of Sciences*, 109(34):13526–13531.
- Diana Rieger, Josephine B Schmitt, and Lena Frischlich. 2018. Hate and counter-voices in the internet: Introduction to the special issue. *SCM Studies in Communication and Media*, 7(4):459–472.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.
- Daniel Russo, Shane Peter Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023a. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023b. [Benchmarking the Generation of Fact Checking Explanations](#). *Transactions of the Association for Computational Linguistics*, 11:1250–1264.
- Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. [Prevalence and psychological effects of hateful speech in online college communities](#). In *Proceedings of the 10th ACM Conference on Web Science, WebSci ’19*, page 255–264, New York, NY, USA. Association for Computing Machinery.
- Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. [Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5157–5163. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th*

- annual meeting of the association for computational linguistics*, pages 1668–1678.
- Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ICA Annual Conference, at Fukuoka, Japan*, pages 1–23.
- Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 39. Cambridge: Cambridge University Press.
- Sandeep Shah, Xiaohong Yuan, and Zanetta Tyler. 2022. An analysis of covid-19 related twitter data for asian hate speech using machine learning algorithms. In *2022 1st International Conference on AI in Cybersecurity (ICAIC)*, pages 1–6. IEEE.
- Alexandra A Siegel. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88.
- Dominik Stammach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.
- Serra Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114.
- Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. [Using pre-trained language models for producing counter narratives against hate speech: a comparative study](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.
- Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. [Generating counter narratives against online hate speech: Data and strategies](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.
- Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. Safer dialogues: Taking feedback gracefully after conversational safety failures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6462–6481.
- Maria Estrella Vallecillo-Rodríguez, Arturo Montejó-Raéz, and Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71:227–245.
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.
- Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting east asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019a. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019b. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the third workshop on abusive language online*, pages 80–93.
- Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303.
- Sharon M. Walter. 1998. [Book reviews: Evaluating natural language processing systems: An analysis and review](#). *Computational Linguistics*, 24(2).
- Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.
- Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.
- Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968.
- Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. [Hate speech and counter speech detection: Conversational context does matter](#). In *Proceedings of the*



2022 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. **What makes good counterspeech? A comparison of generation approaches and evaluation metrics.** In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. **Generate, prune, select: A pipeline for counterspeech generation against online hate speech.** In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

## A Appendix

### A.1 Methodology of the review

Figure 1 shows how the number of published counterspeech papers is subject to a steady growth. To select the studies reviewed in this survey, we follow the PRISMA framework (Moher et al., 2010). The main aim of this review is to provide a guide for tackling counterspeech tasks in the NLP area. Therefore, as inclusion criteria, we only include publicly available papers presenting (i) a computational approach to (ii) text-based tasks that are (iii) related to online counterspeech. In particular, all the included papers either present a data collection, or concern classification, generation or selection tasks. Following previous reviews on counterspeech and abusive language (Vidgen and Derczynski, 2020; Chung et al., 2023), we used three different sources to select the publications of our interest: the ACL Anthology, Scopus and arXiv. First, we searched in these databases for all the publications including at least one of the following keywords: *counterspeech*, *counter-speech*, *counter speech*, *counter narratives*, *counter-narratives*, *counter hate*, *counter-hate*, *counterhate*, *hate countering*, *countering online hate speech*. Similarly to Vidgen and Derczynski (2020), since Scopus includes a much broader content, we limited the subject area to *Computer Science*. The automatic selection resulted in 156 papers from Scopus, 31 from arXiv and 20 from the ACL anthology<sup>14</sup>. 23 duplicates were removed. Then, two of the authors manually revised the automatically filtered publications: first they considered only those which were NLP-related (shown as *NLP & CS* in Figure 1). Then, from this subset, they kept only the publications that either presented a data collection, generation, classification or selection task. They also removed too short (e.g. 2 pages initial studies) or not publicly available papers. The disagreements between the authors (regarding 3 papers) were solved by discussion. After this filtering, a total of 43 papers were included in this survey.

### A.2 Examples of counterspeech taxonomies

As mentioned in Section 2.2, the most widely employed counterspeech taxonomy is the one proposed by Benesch et al. (2016b). However, Qian et al. (2019) make a different distinction, based on the observed strategies adopted by the crowdworkers in their study. These consist of *Identifying*

<sup>14</sup>The search was last implemented on 14 December 2023.

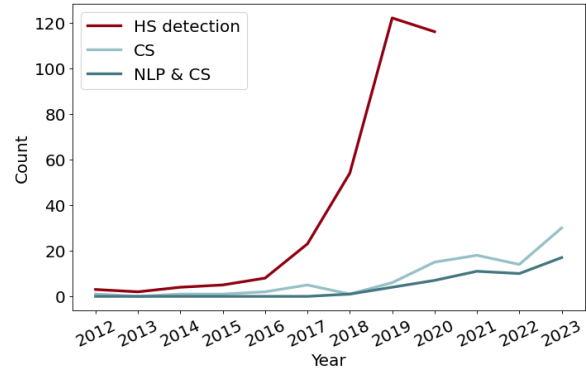


Figure 1: Number of published papers about hate speech detection, counterspeech in general and in the field of NLP. Data for hate speech detection cover only the 2013-2020 time span (Jahan and Oussalah, 2023).

*Hate Keywords*, *Categorize Hate Speech*, *Positive Tone Followed by Transitions*, and *Suggest Proper Actions*. In particular, the strategy of *Identifying Hate Keywords* is based on exhorting users to stop using inappropriate terms. *Categorize Hate Speech* involves the classification of the hate speech into a specific category. *Positive Tone Followed by Transitions* relies on showing empathy first and then proceeding to condemn the hateful text. Finally, with *Suggest Proper Actions* a proactive suggestion is made to the user.

Alternatively, Vidgen et al. (2020) propose a taxonomy where counterspeech are distinguished according to whether it *Rejects the premise of abuse*, *Describes content as hateful or prejudicial*, or *Expresses solidarity with target entities*. Examples of all the identified strategies are shown in Table 3.

### A.3 Datasets for counterspeech-related tasks

In Table 4, we make a non-exhaustive list of available datasets for the tasks described in Section 2.3.

### A.4 Annotators training procedure

Recognizing, post-editing and writing counterspeech requires expertise and practice. When annotators do not have any previous experience in counterspeech they can be trained to acquire proficiency in the task of interest (Chung et al., 2020; Vidgen et al., 2020, 2021; Fanton et al., 2021; He et al., 2021; Furman et al., 2022; Bonaldi et al., 2022b; Gupta et al., 2023; Bonaldi et al., 2023). The most employed procedure for annotators' training includes the following steps:

- reading and discussing NGO guidelines and public documentation describing the activity

	Strategy	Example
Benesch et al. (2016b)	<b>Presenting facts</b>	Actually homosexuality is natural. Nearly all known species of animal have their gay communities. Whether it be a lion or a whale, they have or had (if they are endangered) a gay community.
	<b>Pointing out hypocrisy</b>	The 'US Pastor' can't accept gays because the Bible says not to be gay. But...he ignores: The thing about eating shrimp or pork, [...] The thing about working on the Holy Day (Saturday or Sunday depending)...for any and all of those sins one should burn for an eternity, yet is ignored.
	<b>Warning of consequences</b>	I'm not gay but nevertheless, whether You are beating up someone gay or straight, it is still an assault and by all means, this preacher should be arrested for sexual harassment and instigating!!!
	<b>Affiliation</b>	Hey I'm Christian and I'm gay and this guy is so wrong. Stop the justification and start the accepting. I know who my heart and soul belong to and that's with God: creator of heaven and earth. We all live in his plane of consciousness so it's time we started accepting one another. That's all.
	<b>Denouncing hateful speech</b>	please take this down YouTube. this is hate speech.
	<b>Humor and sarcasm</b>	Of course Jews are focused on 'world domination', even "galaxy domination". But so are Sith Order, Sauron etc.
Mathew et al. (2019)	<b>Positive tone</b>	I am a Christian, and I believe we're to love everyone!! No matter age, race, religion, sex, size, disorder...whatever!! I LOVE PEOPLE!! We are not going to go anywhere as a country if we don't put God first in our lives, and treat EVERYONE with respect.
	<b>Hostile language</b>	This is ridiculous!!!!!! I hate racist people!!!! Those police are a**holes!!!
Chung et al. (2019)	<b>Counter-questions</b>	Is this true? Where is your source?
Qian et al. (2019)	<b>Identify Hate Keywords</b>	The C word and language attacking gender is unacceptable. Please refrain from future use.
	<b>Categorize Hate Speech</b>	The term fa**ot comprises homophobic hate, and as such is not permitted here.
	<b>Positive Tone Followed by Transitions</b>	I understand your frustration, but the term you have used is offensive towards the disabled community. Please be more aware of your words.
	<b>Suggest Proper Actions</b>	I think that you should do more research on how resources are allocated in this country.
Vidgen et al. (2020)	<b>Reject the premise of abuse</b>	it isn't right to blame China!
	<b>Describe content as hateful or prejudicial</b>	you shouldn't say that, it's derogatory
	<b>Express solidarity with target entities</b>	Stand with Chinatown against racists.

Table 3: Taxonomies of counterspeech proposed by various authors. Both Mathew et al. (2019) and Chung et al. (2019) add new categories to the classes proposed by Benesch et al. (2016b). All the reported examples come from the relative papers, except for the *Humor and sarcasm* example, which is taken from Fanton et al. (2021) dataset.

Dataset	Task	Size	Interact.	Coll.	Source
Lee et al. (2022b)	Counter-trolling	6,686	Pairs	Crawl. Crowd.	and Reddit
Chakravarthi (2020)	Hope speech	59,354	Single c.	Crawl.	YouTube
García-Baena et al. (2023)	Hope speech	1,650	Single c.	Crawl.	Twitter
Palakodety et al. (2019)	Hope speech	921,235	Single c.	Crawl.	YouTube
Kim et al. (2022b)	Prosocial dialogue	58,137	Dialog.	Hybr.	Morality-related data
Logacheva et al. (2022)	Detoxification	12,000	Pairs	Crowd.	Toxic sentences data
Ung et al. (2022)	Feedback on safety failures	7,881	Dialog.	Hybr.	Xu et al. (2021)
Stammach and Ash (2020)	Misinformation countering	67,687	Triplets	Hybr.	Thorne et al. (2018)
Kotonya and Toni (2020)	Misinformation countering	11,832	Pairs	Crawl.	Fact-checking websites
Ma et al. (2023)	Misinformation countering	690,047	Pairs	Crawl.	Twitter
Alhindi et al. (2018)	Misinformation countering	12,836	Triplets	Crawl.	PolitiFact
Russo et al. (2023b)	Misinformation countering	8,289	Triplets	Crawl.	Fact-checking websites
Russo et al. (2023a)	Misinformation countering	11,990	Triplets	Hybr.	Full Fact

Table 4: Available datasets on tasks related to counterspeech. The data collected by Russo et al. (2023b) are not distributed, but they share the code to replicate the data collection.

of counterspeech writing;

- b) reading both examples of counterspeech writing and of the specific task of interest (e.g. post-editing) performed by experts;
- c) practice the task on a subsample of examples;
- d) discuss disagreements with an expert.

This procedure can last from two to four weeks. Table 5 summarises the steps undertaken by the studies explicitly describing how they trained the annotators. Furthermore, it is important to preserve the *well-being* of the annotators, given the risks involved in working with hateful content. In particular, taking simple precautions like those suggested by Vidgen et al. (2019b) is enough to safeguard the annotators’ mental health. These precautions include explaining the prosocial purpose of the research, limiting the annotation time to no more than three hours per day, taking regular breaks, and having several meetings to allow for possible problems or distress to emerge.

Study	a	b	c	d
Chung et al. (2020)	✓	✓	-	-
He et al. (2021)	-	-	✓	✓
Vidgen et al. (2021)	-	-	-	✓
Fanton et al. (2021)	✓	✓	✓	✓
Bonaldi et al. (2022b)	✓	✓	✓	✓
Gupta et al. (2023)	✓	-	-	-
Bonaldi et al. (2023)	-	✓	-	✓

Table 5: The steps for annotators’ training in the studies that explicitly mention them, as described in §A.4.